# Wiki정의로부터 ISA를 추출할 수 있는 언어적 규칙

한영석*, 오창근**
*수원대학교 정보미디어학과
**수원대학교 컴퓨터학과
e-mail:ckdrms@suwon.ac.kr

## An Automatic Construction of ISA relations of Wordnet Using Wiki Definitions

Yeong-suk Han*, Chang-guen Oh**
*Dept of Information Media, Suwon University
**Dept of Computer Engineering, Suwon University

요 약

워드넷(WordNet)의 논리적 내포관계(ISA)를 자동으로 WIKI와 같은 동적인 백과사전으로부터 구할 수 있다면, 워드넷과 같은 지식베이스를 전문분야로까지 쉽게 확장할 수 있을 것이다. 또한 동적인 백과사전에 기반하기 때문에 지식베이스의 동적인 업데이트가 가능하게 된다. 본 논문은 워드넷과 같은 정적이고 수동으로 제작된 개념망이 온라인상의 동적 백과사전에 의해서 어느 정도 자동화 될 수 있는지 밝히고자 하였다. 워드넷의 IT관련 100개의 표제어에 대해서 WIKI 백과사전에서 추출한 정의를 이용하여 ISA관계를 구축하고 그 결과가 워드넷과 어느 정도 일치하는지를 실험하였다. 실험결과 자동구축된 ISA관계는 워드넷에 대하여 80%의 일치율을 보였다.

ABSTRACT

The paper aims at showing the subsumption relations of the Wordnet can be captured automatically from a dynamic encyclopedia such as Wikipedia with a meaningful precision. The idea behind the proposal is that a knowledge base in the form of Wordnet can be dynamically obtained and maintained accordingly to the online dictionaries so that the scalability of knowledge base construction may be achieved to some degree. To show the plausibility of dynamic ISA construction, we have tested how well the ISA relations among the 100 technology terms selected from the Wordnet can be saved from the ISA construction by the wiki definitions of the selected terms. As a result the wiki definition led to the ISA relations of the Wordnet with the precision of 80%.

## 1. 서론

워드넷(WordNet)[1]과 같은 ISA는 정보검색과 자연어처리 등에서 유용하게 사용될 수 있으나[2], 용어의 커버리지가 제한적이고 새롭게 등장하는 관심용어를 제때 반영하지 못함으로 응용의 한계를 가지고 있다.

현재 워드넷(WordNet)과 같이 단어, 문장의 의미를 가지고 그들 사이의 내포관계를 어휘목록으로 작성하기 위해서는 많은 시간과 비용을 소비하여야 한다.[2] 어휘에 대한 평가를 내리고 어휘간의 관련성을 연구 파악하는 것은 쉽지 않은 일이다. 수동작업은 정확성은 높일 수 있으나 일관성에 문제가 있을 수 있으며, 어휘선정에 있어서 주관이 개입될 수 있다. 새로운 어휘가 계속 만들어지고 있다. 특히 시소러스의 하위에 위치하는 고유명사 및 전문용어들은 끊임없이 만들어 지고 있지만, 이들을 시소러스와 같은 지식구조에 수동으로 반영하지 못하고 있다. 다행히 사전적 정의는 계속 진화하고 있음에 착안하여 본 논문에서는 위에서 제기한 문제점들을 해결하고자 워드넷이 담고 있는 ISA관계를 자동으로 동적인 사전에서 추출할 수 있는 방법을 제안한다. 더불어 동적인 사전의 품질의 한계에도 불구하고, 동적 ISA정보 베이스를 구축하는데 어느 정도 유용한지를 밝히고자 한다.

워드넷(WordNet)에 포함된 일부 단어들을 가지고 대표적인 동적인 사전인 위키백과사전(Wiki-pedia)의 정의표현을 추출한 후, 정의표현을 가공하여 용어간의 ISA관계를 구축하고, 원래의 워드넷 ISA정보와 비교하였다. 정의표현으로부터 ISA관계를 추출하는 데는 정의문의 특성에 구문적 특성을 반영하는 정규문법(regular grammar)수준의 규칙을 이용하여 파싱하였다. 실험결과 위키사전의 정의문으로부터 80%의 워드넷 ISA관계를 추출할 수 있었다.

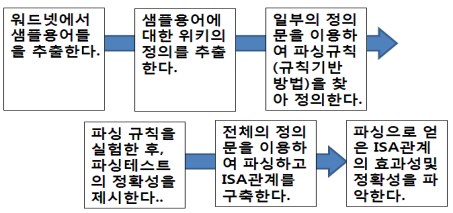
## 2. 관련연구

기존의 여러 연구들이 코퍼스나 기타의 언어자원을 활용한 어휘자원의 자동구축에 초점을 맞추었다.[3] 여러 연

구들에서 공통으로 목표한 것은 데이터를 분류하고 구축하는데 많은 노력과 시간이 걸리는 이른바 "지식습득의 병목현상(knowledge acquisition bottleneck)"의 문제를 해결하기 위한 방식으로 더 많은 데이터를 더 정확하게 기계적으로 분석하기 위한 방법론들이다. 하지만 이러한 기존의 여러 연구들에서는 정적인 환경의 백과사전 속 언어자원을 사용하고 있다. 끊임없이 진화하는 언어자원의 반영이 이루어지 않은 문제점을 가지고 있다는 것은 새로 습득한 지식에 관하여서 또 하나의 병목현상이 이루어지고 있다는 것이다. 본 논문에서는 이러한 문제점을 착안하여 동적인 환경의 사전 속 언어자원을 활용하고 있으며, 워드넷의 상하위어 관계를 활용해서 영어 텍스트에서 발견되는 어휘들 간에 계층적 개념관계를 자동으로 구축하는 방법론을 제시하고[5], 동사의 ISA를 반자동으로 구축하고자, 정확한 규칙을 사용하여서 사전의 정의문에서 중심어를 추출하는 방식을 제시하고[6] 있는 연구들에서는 추출된 개념관계(상/하위어)가 정확한지에 대한 검증은 결여되어 있다. 반면에 사전의 정의문에서 발견되는 일정한 패턴을 활용해서 상위어를 추출하고 추출된 상하위어의 정확도를 화자 직관을 활용하여 검증하는[3] 연구들은 화자 직관을 동원한 상하위어의 판별이 체계적인 의미 관계를 규정한 하나의 틀 안에서 일관적으로 성립되었는지는 알 수 없다. 화자 직관은 개인별이나 집단별로 차이가 있기 때문에, 체계적으로 성립한 하나의 의미체계라는 틀에서 검증되었는지는 담보될 수 없다. 그러므로 기존 연구에서 활용한 화자 직관을 통한 판별 작업이 전체 실험을 통해서 살펴볼 때, 일관적인 상하위어 관계성을 의미하는지를 알 수 없다. 본 논문에서는 이러한 문제점을 착안하여 기초 원천 자원과의 비교를 통한 검증법을 택하고 있다.

## 3. 동적사전을 이용한 파싱규칙 추출

워드넷의 경우, 많은 시간과 비용을 투자하여 언어들간의 규칙을 찾고 용어간의 관계를 정의하였지만 동적으로 진화하는 사전환경에서는 그 실효성이 부족하다. 본 논문에서는 동적으로 진화하는 위키사전 속에서 ISA관계의 파악을 위한 규칙기반의 방법을 제안하고, 이를 통한 시소러스 자동구축을 전망함으로써 고비용을 요구하고 경직된 정보를 담은 워드넷의 문제에 대한 대안을 제안하고자 한다. 연구의 흐름은 <그림1>과 같다.

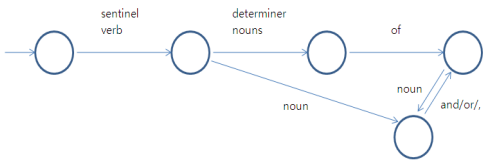



<그림1> 연구흐름

워드넷(WordNet)에서 검색결과로 나타나는 단어들을 가지고 최근 들어 백과사전으로 각광받고 있는 위키백과사전(Wiki-pedia)에서 찾아 검색한 뒤, 워드넷과 같은 구조의 표본그룹을 만든다. 그 중 일부를 가지고 파싱모델을 구축하고 나머지 일부를 가지도 파싱모델을 검증한다.

표본그룹에서 정의부분을 찾아 그 속에서 나타나는 내포관계(ISA)의 패턴을 찾는다. 최종적으로 'A는B다'의 형태로써 'A'라는 단어가 'B'라는 정의명사로 정의되어짐으로써 상/하위 관계를 갖는 것과 같은 네트워크 구조를 찾는 것을 목표로 한다.

워드넷(WordNet)에서 나타나는 전산관련 용어 100개를 대상으로 하여, 위키백과사전(Wiki-pedia)에서 워드넷과 같은 형태의 구조로 찾아낸 표본그룹에 파싱규칙을 발견하여 정리 제안한다.<그림2>




<그림2> 파싱규칙

파싱규칙에서 사용되는 용어를 정리한다.

[s : Sentinel verb]

내포관계(ISA)를 표현하기 위해 나타나는 특정 동사이다. 'be동사'와 'refer to'의 형태를 찾아 볼 수 있으며, 정의 표현이 확실하지 않은 문장에서는 특정 동사를 찾아볼 수 없었다.

| s : sentinel verb | |
|---|---|
| 내용 | be 동사 : is, are |
| | (may) refer to : refer to |
| | - : 특정 동사 없음 |

[d of : determiner nouns]

정의명사를 직접적으로 표현함을 벗어나 한정사를 동반하는 형태를 가진다.

| d of : (deteminer nouns) of | | |
|---|---|---|
| 내용 | a series of | subfield of |
| | collection of | set of |
| | abbreviation of | type of |
| | piece of | part of |

[a/o : and / or]

정의명사가 하나이상의 형태를 가진다. 상위어의 세분화를 볼 수 있다.

| a/o : and / or | |
|---|---|
| 내용 | and / or |

[규칙 1.] Ⓐ s. Ⓑ

'Ⓐ는 Ⓑ이다'의 전형적인 내포관계(ISA)를 보여주는 형태이다.

ex) A Computer is a machine that manipulates data according to a list of instructions.

->'computer는 machine이다'

[규칙 2.] Ⓐ s. d of. Ⓑ

Ⓐ와 Ⓑ 사이의 관계가 한정사를 동반하여, 'Ⓐ는 Ⓑ의 일부이다' , 'Ⓐ는 Ⓑ들로 구성된다' 와 같은 형태이다.

ex)A database is a structured of records or data.

->'database는 record들로 구성된다'

[규칙 3.] Ⓐ s. Ⓑ Ⓑ

'Ⓐ는 Ⓑ이다'의 확장적인 형태로써 Ⓐ를 Ⓑ라는 단순 정의명사로 정의할 수 없어 ⒷⒷ 형태의 복수명사를 가지는 형태이다.

ex) File Transfer Protocol(FTP) is network protocol used to transfer data from one computer to another through a network, such as the Internet.

-> 'FTP는 network protocol이다'

[규칙 4.] Ⓐ s. Ⓑ a/o. Ⓑ

'Ⓐ는 Ⓑ이다'의 확장적인 형태로써 Ⓐ를 여러 Ⓑ라는 정의명사를 통하여 정의 할 수 있는 경우이다.

Ⓑ and/or Ⓑ 형태의 복수 정의를 볼 수 있다.

ex) Homepage is the URL or local file that automatically loads when a web brower starts and when the browser's "home"button is pressed.

-> 'Homepage는 URL 또는 local file이다'

[규칙 5.] Ⓐ -

Ⓐ와 Ⓑ사이의 관계가 정의 형태가 아닌 Ⓐ에서 나타나는 현상을 문장으로 정의한 형태로써 특정한 정의 명사를 찾아 볼 수 없다.

ex) Logging, the method whereby a user obtains access to a computer system.

-> 로그인을 현상묘사로써 설명한다.

제안한 5개의 패턴에 관하여 표본대상에서 보여지는 빈도를 <표1>로 나타내었다.

| 정의형태 | 빈도(%) |
|---|---|
| Ⓐ s Ⓑ | 40% |
| Ⓐ s d of Ⓑ | 16% |
| Ⓐ s Ⓑ Ⓑ | 24% |
| Ⓐ s Ⓑ a/o Ⓑ | 12% |
| Ⓐ - | 8% |

<표1> 표본그룹 실험

## 4. 실험

워드넷(WordNet)에서 나타나는 전산관련 용어 50개를 대상으로 하여, 위키백과사전(Wiki-pedia)에서 워드넷과 같은 형태의 구조로 찾아낸 실험그룹에 3장에서 제안한 파싱규칙을 적용하여 보여지는 빈도를 <표2>로 나타내었다.

| 정의형태 | 빈도(%) |
|---|---|
| Ⓐ s Ⓑ | 40% |
| Ⓐ s d of Ⓑ | 16% |
| Ⓐ s Ⓑ Ⓑ | 16% |
| Ⓐ s Ⓑ a/o Ⓑ | 16% |
| Ⓐ - | 12% |

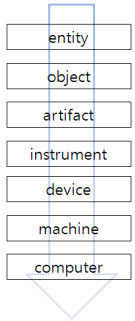
<표2> 실험그룹 실험

표본그룹에서 보여지는 파싱규칙을 가지고 실험그룹에 적용하여 실험한 결과 96%의 파싱규칙 정확성을 보여주어, 제시된 파싱규칙을 ISA관계 복원을 위한 규칙기반 방법으로 제시하였다.<표3>

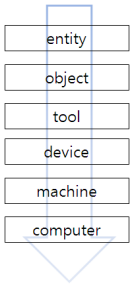
| | | | |
|---|---|---|---|
| ISA추출 (규칙1~규칙4) | 44개 | 성공 | 44개 |
| | | 실패 | 0개 |
| - (규칙5) | 6개 | 성공 | 5개 |
| | | 실패 | 1개 |

<표3> 파싱규칙 정확성

본 연구는 기존에 이루어지고 있는 연구에서 나타나는 정적인 환경의 사전에서 단어 정의간의 의미관계를 파악하여 시소러스의 자동구축을 하는 것은 기초토대로 하였다. 더불어 동적으로 진화하고 있는 사전을 가지고 워드넷을 작성하는 것이 얼마나 유효한지를 보여주기 위하여 전체표본을 대상으로 내포구조<그림3><그림4>의 비교를 통하여 보인다.




<그림3> 워드넷(word-net) 내포구조
computer에 대한 검색결과




<그림4> 위키백과사전(wiki-pedia) 내포구조
computer에 대한 검색결과

본 논문에서 제시하는 파싱규칙을 통해 복원된 ISA관계와 워드넷의 검색결과로 나타나는 단어사이의 일치성과 트리구조의 형태를 비교한 결과 80%의 정확성을 보였으며, 정확성을 벗어나는 범위에서는 워드넷에서 보여지지 않았던 새로운 링크가 포함되어 있었다.

## 5. 결론

워드넷은 많은 비용과 시간을 들여 구축하였으나, 새로운 용어를 추가하는 등의 지속적인 관리를 해야 하고, 특히 전문적이고 빠르게 변화하는 기술이나 전문분야의 용어들은 다 커버할 수 없는 한계점을 가지고 있다. 이전 연구에서 정적인 백과사전을 이용하여 ISA관계를 추출하는 시도들이 있었으나, 용어들이 제한적이여서 동적인 ISA 지식베이스를 구축하기에는 적절하지 않다.

워드넷의 100개의 용어를 대상으로 Wiki의 정의표현을 추출하고, 본 논문에서 설계한 정의문 파싱규칙을 이용하여 ISA관계를 추출하였다. 그 결과 80%의 워드넷 ISA관계를 복원할 수 있었다. 파싱규칙을 정교하게 함으로써 정확성을 높일 수 있다. 본 논문의 요점은 품질에 대한 논란에도 불구하고 Wiki와 같은 동적인 백과사전을 통하여 워드넷과 같은 수동구축 ISA의 관계를 복원함으로써 동적인 지식베이스를 구축할 수 있는 가능성을 제시하는데 있다.

## 참고문헌

[1] G.A.Miller, "WordNet: An On-Line Lexical Database", International Journal of lexicography, 1990.
[2] Wikipedia, http://en.wikipedia.org/wiki/WordNet
[3] 김민수,김태연,노봉남 "국어사전을 이용한 한국어 명사에 대한 상위어 자동 추출 및 WordNet의 프로토타입 개발" 한국정보처리학회, 1995.
[4] 신명근, "Concept Hierachy Creation Using Hypernym Relatonship" 한국컴퓨터정보학회, 2006.
[5] 김혜경,윤애선, "동사 어휘의미망의 반자동 구축을 위한 사전정의문의 중심어 추출",언어와정보,2006